

# Efficient acquisition rules for model-based approximate Bayesian computation

Marko Järvenpää\*, Michael U. Gutmann†, Aki Vehtari\* and Pekka Marttinen\*

\*Helsinki Institute for Information Technology HIIT, Department of Computer Science, Aalto University,

†School of Informatics, University of Edinburgh

## Abstract

Approximate Bayesian computation (ABC) is a method for Bayesian inference when the likelihood is unavailable but simulating from the model is possible. However, many ABC algorithms require a large number of simulations, which can be costly. To reduce the computational cost, surrogate models and Bayesian optimisation (BO) have been proposed. Bayesian optimisation enables one to intelligently decide where to evaluate the model next, but standard BO strategies are designed for optimisation and not specifically for ABC inference. Our paper addresses this gap in the literature. We propose a new acquisition rule that selects the next evaluation where the uncertainty in the posterior distribution is largest. Experiments show that the proposed method often produces the most accurate approximations, especially in high-dimensional cases or in the presence of strong prior information, compared to common alternatives.

## 1 INTRODUCTION

We consider the problem of Bayesian inference of some unknown parameters  $\theta \in \Theta \subset \mathbb{R}^p$  of a simulation model. Such models are typically not amenable to any analytical treatment but they can be simulated with any parameter  $\theta \in \Theta$  to produce data  $\mathbf{x}_\theta \in \mathcal{X}$ . Simulation models are also called simulator-based or implicit models [Diggle and Gratton, 1984]. Our prior knowledge about the unknown parameters  $\theta$  is represented by the prior probability density  $\pi(\theta)$  and the goal of the analysis is to update our knowledge about the parameters  $\theta$  after we have observed data  $\mathbf{x}_{obs} \in \mathcal{X}$ .

If evaluating the likelihood function  $\pi(\mathbf{x} | \theta)$  is feasible,

the posterior distribution can be computed from Bayes' theorem

$$\begin{aligned} \pi(\theta | \mathbf{x}_{obs}) &= \frac{\pi(\theta)\pi(\mathbf{x}_{obs} | \theta)}{\int_{\Theta} \pi(\theta')\pi(\mathbf{x}_{obs} | \theta') d\theta'} \\ &\propto \pi(\theta)\pi(\mathbf{x}_{obs} | \theta). \end{aligned} \quad (1)$$

However, if one can only simulate from the model, that is, draw samples  $x_\theta \sim \pi(\cdot | \theta)$ , and not evaluate the likelihood function  $\pi(\mathbf{x}_{obs} | \theta)$ , the standard Bayesian approach cannot be used.

Approximate Bayesian computation (ABC) replaces likelihood evaluations with model simulations<sup>1</sup>, see e.g. [Marin et al., 2012, Turner and Van Zandt, 2012, Lintusaari et al., 2016] for an overview. The main idea of the basic ABC algorithm is to draw a parameter value from the prior distribution, simulate a data set with the given parameter value, and accept the value as a draw from the posterior if the discrepancy between the simulated and observed data is small enough. This algorithm produces samples from the approximate posterior distribution

$$\pi_{ABC}(\theta | \mathbf{x}_{obs}) \propto \pi(\theta) \int \pi_\varepsilon(\mathbf{x}_{obs} | \mathbf{x})\pi(\mathbf{x} | \theta) d\mathbf{x}, \quad (2)$$

where  $\pi_\varepsilon(\mathbf{x}_{obs} | \mathbf{x}) \propto \mathbb{1}_{\Delta(\mathbf{x}_{obs}, \mathbf{x}) \leq \varepsilon}$ , although other choices of  $\pi_\varepsilon$  are also possible [Wilkinson, 2013]. The function  $\Delta : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$  is the discrepancy that tells how different the simulated and observed data sets are, and it is often formed by combining a set of summary statistics. Choosing an appropriate discrepancy is an active research topic [Fearnhead and Prangle, 2012, Blum et al., 2013], but in this article we assume a suitable discrepancy is provided to us. The threshold  $\varepsilon$  controls the trade-off between the accuracy of the approximation and computational cost: a small  $\varepsilon$  yields accurate approximations but requires more simulations, see e.g. [Marin et al., 2012]. Given  $t$  samples from the model for some  $\theta$ , that

<sup>1</sup>Such approaches are also called likelihood-free inference.

is,  $\mathbf{x}_\theta^{(i)} \sim \pi(\cdot | \theta)$  for  $i = 1, \dots, t$ , the value of the ABC posterior in Equation (2) can be estimated as

$$\pi_{\text{ABC}}(\theta | \mathbf{x}_{\text{obs}}) \propto \pi(\theta) \sum_{i=1}^t \pi_\varepsilon(\mathbf{x}_{\text{obs}} | \mathbf{x}_\theta^{(i)}). \quad (3)$$

Algorithms based on MCMC and sequential Monte Carlo have been used to improve the efficiency of ABC [Beaumont et al., 2009, Toni et al., 2009, Marin et al., 2012, Lenormand et al., 2013]. Unfortunately, the sampling based methods still require a very large number of simulations. In this paper we focus on the challenging scenario where the number of available simulations is very limited, e.g. to fewer than a thousand, rendering standard ABC methods infeasible.

Different modelling approaches have been proposed to reduce the number of simulations required. For example, in the synthetic likelihood method summary statistics are assumed to follow the Gaussian density [Wood, 2010, Price et al., 2016] and the resulting likelihood approximation can be used together with MCMC. Wilkinson [2014], Meeds and Welling [2014], Jabot et al. [2014], Kandasamy et al. [2015], Drovandi et al. [2015], Gutmann and Corander [2016], Järvenpää et al. [2016] all use Gaussian processes (GP) to accelerate ABC in various ways. Some other alternative approaches are considered by Fan et al. [2013], Papamakarios and Murray [2016].

While probabilistic modelling has been used to accelerate ABC inference, and strategies have been proposed for selecting which parameter to simulate next, little work has focused on trying to quantify the amount of uncertainty in the estimator of the ABC posterior density itself. This uncertainty is due to finite computational budget to perform the inference. Consequently, little has been done to design strategies that directly aim to minimise this uncertainty. To our knowledge, only Kandasamy et al. [2015] have used the uncertainty in the likelihood function to propose new simulation locations. However, they assumed that the likelihood can be evaluated, although with high computational cost. Also, Wilkinson [2014] modelled the uncertainty in the likelihood to rule out regions with negligible posterior probability. Rasmussen [2003] used GP regression to accelerate Hybrid Monte Carlo but did not consider ABC. Finally, Gutmann and Corander [2016] proposed Bayesian optimisation (BO) to efficiently select new evaluation locations, but the BO strategies they used to illustrate the framework are originally designed for optimisation and not for ABC.

In this article we propose an acquisition function tailored specifically for ABC. The acquisition function measures the uncertainty in our current estimate of the posterior

density function, and proposes the next point to simulate where this uncertainty is maximal. In Section 2 we formulate our approach on a general level. In section 3 we propose a particular algorithm, based on modelling the discrepancy with a GP. Section 4 contains experiments. Some additional details are discussed in Section 5 and Section 6 concludes.

## 2 PROBLEM FORMULATION

We start by presenting the main idea of the proposed framework. Suppose we have training data  $D_t = \{(\mathbf{x}_i, \theta_i)\}_{i=1}^t$  and our uncertainty about some future model output  $\mathbf{x}^* \in \mathcal{X}$  with parameter  $\theta^*$  is represented by a probability measure  $\pi(\mathbf{x}^* | \theta^*, D_t)$ <sup>2</sup>. Now, our estimate  $\pi_{\text{ABC}}$  for the ABC posterior density depends on the training data according to Equation (3), and is therefore also a random quantity. We represent the uncertainty in  $\pi_{\text{ABC}}$  using a probability measure  $\pi(\pi_{\text{ABC}} | D_t)$  over the space of (suitable smooth) density functions  $\pi_{\text{ABC}} : \Theta \rightarrow \mathbb{R}_+$ .

Let  $\mathcal{L}(\pi(\pi_{\text{ABC}}))$  denote the loss due to our uncertain knowledge of the ABC-posterior density. This loss function could measure, for example, the overall uncertainty in the density  $\pi_{\text{ABC}}$ , the uncertainty in a particular point estimate of interest, or the maximum uncertainty over all possible parameter values. If  $m$  future simulations can be performed, our aim is to choose such evaluation locations  $\theta^* = \{\theta_{t+1}, \theta_{t+2}, \dots, \theta_{t+m}\}$  that the expected loss after simulating the model at these locations is minimised, that is, we want to minimise

$$\begin{aligned} & \mathbb{E}_{\mathbf{x}^* | \theta^*, D_t}(\mathcal{L}(\pi(\pi_{\text{ABC}} | \mathbf{x}^*, \theta^*, D_t))) \\ &= \int \mathcal{L}(\pi(\pi_{\text{ABC}} | \mathbf{x}^*, \theta^*, D_t)) \pi(\mathbf{x}^* | \theta^*, D_t) d\mathbf{x}^*, \quad (4) \end{aligned}$$

where we average over the simulator outputs  $\mathbf{x}^* = \{\mathbf{x}_{t+1}, \dots, \mathbf{x}_{t+m}\}$  at parameters  $\theta^*$ .

This approach resembles the entropy search method [Hennig and Schuler, 2012, Hernández-Lobato et al., 2014]. Other related approaches have been proposed by Wang et al. [2016], Bijl et al. [2016]. Different from these approaches, our main goal is to select the parameter for a future run of the costly simulation model so that the uncertainty in the approximate posterior is minimised. Entropy search, in contrast, aims to find a parameter value that maximises the objective function, and minimise the uncertainty related to this maximiser.

The framework outlined above requires some modelling choices and leads to computational challenges (as is the

<sup>2</sup>We assume that the set  $\mathcal{X}$  is such that this can be well-defined. Alternatively, we could model some summary statistics of the full data  $\mathbf{x}$  or the discrepancy.

case with the entropy search). Therefore, in the next section we propose an approximate but tractable strategy. Furthermore, we restrict our discussion to a sequential setting, i.e.  $m = 1$ , and leave extensions to multiple simultaneous acquisitions for future work.

Details of our approach appear in the next section, but the main idea is shown in Figure 1. We model the discrepancy  $\Delta_\theta = \Delta(\mathbf{x}_{obs}, \mathbf{x}_\theta)$  with GP regression (Fig. 1b). The ABC posterior is proportional to the prior times the probability of obtaining a discrepancy realisation that is below the threshold when the model is simulated. However, because the GP is fitted with limited training data, this probability cannot be estimated exactly, causing uncertainty in the (approximate) posterior density function (Figure 1a). We propose an acquisition function that selects as the next point to simulate the location where this uncertainty is greatest.

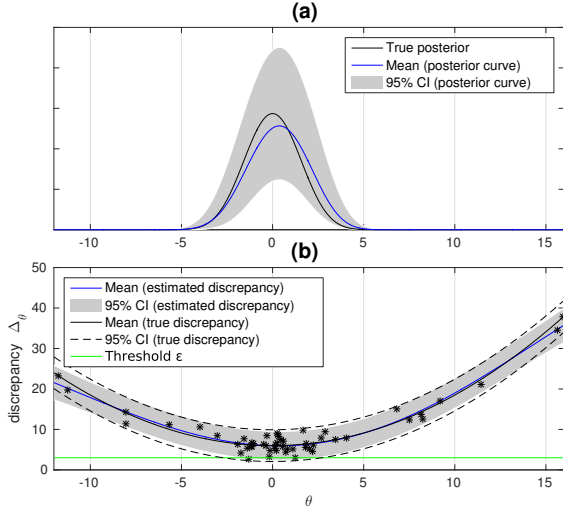


Figure 1: (a) The blue curve shows the mean of the estimated posterior density function and the grey area its 95% pointwise credible interval. (b) The estimated and true discrepancy distributions are compared. Most of the evaluations are successfully chosen on the modal area of the posterior, leading to a good approximation.

### 3 NONPARAMETRIC MODELLING AND PARAMETER ACQUISITION

Next we present the details of our approach. Section 3.1 describes the GP model for the discrepancy, which permits closed-form equations for the mean and variance of the ABC posterior estimate, derived in Section 3.2. Handling GP hyperparameters is discussed in Section 3.3, and in Section 3.4 we formulate a deterministic and stochastic versions of the proposed acquisition rules.

#### 3.1 GP MODEL FOR THE DISCREPANCY

We model the discrepancy  $\Delta_\theta$ , which is a stochastic process indexed by  $\theta$ , as a function of the parameter  $\theta$ , but some alternatives are briefly outlined in supplementary material. We assume that the discrepancy can be modelled by a Gaussian distribution for each parameter value  $\theta$ , that is  $\Delta_\theta \sim \mathcal{N}(f(\theta), \sigma_n^2)$  for some unknown suitably smooth function  $f : \Theta \rightarrow \mathbb{R}$  and variance  $\sigma_n^2 \in \mathbb{R}_+$  both of which need to be estimated. We place a Gaussian process prior on  $f$  so that  $f \sim \mathcal{GP}(\mu(\theta), k(\theta, \theta'))$ . While other choices are also possible, in this paper we consider  $\mu(\theta) = 0$ , and use the squared exponential covariance function  $k(\theta, \theta') = \sigma_f^2 \exp(-\sum_{i=1}^p (\theta_i - \theta'_i)^2 / (2l_i^2))$ . There are thus  $p + 2$  hyperparameters to estimate, denoted by  $\phi = (\sigma_f^2, l_1, \dots, l_p, \sigma_n^2)$ .

Conditioned on the obtained training data  $D_t = \{(\Delta_i, \theta_i)\}_{i=1}^t$ , which consists of realised discrepancy-parameter pairs, and the GP hyperparameters  $\phi$ , our knowledge of the function  $f$  evaluated at an arbitrary point  $\theta^* \in \Theta$  can be shown to be  $f(\theta^*) | D_t, \theta^*, \phi \sim \mathcal{N}(m(\theta^*), v^2(\theta^*))$ , where

$$m(\theta^*) = k(\theta^*, \theta) K(\theta, \theta)^{-1} \Delta, \quad (5)$$

$$v^2(\theta^*) = k(\theta^*, \theta^*) - k(\theta^*, \theta) K(\theta, \theta)^{-1} k(\theta, \theta^*) \quad (6)$$

and  $K(\theta, \theta) = k(\theta, \theta) + \sigma_n^2 \mathbf{I}$ . Above we have overloaded the notation and defined  $k(\theta^*, \theta) = (k(\theta^*, \theta_1), \dots, k(\theta^*, \theta_t))^T$ ,  $k(\theta, \theta)_{ij} = k(\theta_i, \theta_j)$  for  $i, j = 1, \dots, t$  and similarly for  $k(\theta, \theta^*)$ . We have also used  $\Delta = (\Delta_1, \dots, \Delta_t)^T$ . A comprehensive presentation of GP regression can be found in [Rasmussen and Williams \[2006\]](#).

#### 3.2 QUANTIFYING THE UNCERTAINTY OF THE POSTERIOR ESTIMATE

We first assume that the GP hyperparameters  $\phi$  are known, that the parameter space of the simulation model  $\Theta$  is bounded, and that the prior is uniform on its support  $\Theta$ , that is,  $\pi(\theta^*) = \mathbf{1}_{\theta^* \in \Theta} / |\Theta|$ . We will discuss relaxing these assumptions later.

As in [\[Gutmann and Corander, 2016\]](#), one can compute the posterior predictive density for new discrepancy value at  $\theta^*$  using  $\Delta_{\theta^*} | D_t, \theta^*, \phi \sim \mathcal{N}(m(\theta^*), v^2(\theta^*) + \sigma_n^2)$  and obtain a model-based estimate for the acceptance probability given the modelling assumptions and training data  $D_t$ , as  $\mathbb{P}(\Delta_{\theta^*} \leq \epsilon) = \Phi((\epsilon - m(\theta^*)) / \sqrt{\sigma_n^2 + v^2(\theta^*)})$ , where  $\Phi(z) = \int_{-\infty}^z \exp(-t^2/2) dt / (2\pi)$  is the cdf of the standard normal distribution. This probability is approximately proportional to the likelihood and yields a useful point esti-

mate of the likelihood function. We here take a different approach and explicitly exploit the fact that part of the probability mass of  $\Delta_\theta$  is due to our uncertainty in  $f$ . Indeed, if we knew  $f$ , the acceptance probability would be<sup>3</sup>

$$p(\theta^*) = \Phi\left(\frac{\varepsilon - f(\theta^*)}{\sigma_n}\right). \quad (7)$$

With a limited number of discrepancy-parameter pairs in  $D_t$  there is uncertainty in the values of the function  $f$  (and in GP hyperparameters  $\phi$ ). We can obtain the expectation of the acceptance probability  $p(\theta^*)$  at  $\theta^*$  by using the law of the unconscious statistician, that is

$$\begin{aligned} \mathbb{E}(p(\theta^*)) &= \int_{-\infty}^{\infty} \Phi\left(\frac{\varepsilon - f}{\sigma_n}\right) \mathcal{N}(f | m(\theta^*), v^2(\theta^*)) df \\ &= \Phi\left(\frac{\varepsilon - m(\theta^*)}{\sqrt{\sigma_n^2 + v^2(\theta^*)}}\right), \end{aligned} \quad (8)$$

where  $m(\theta^*)$  and  $v^2(\theta^*)$  are given by Equations (5) and (6). The second equality follows from the fact  $\Phi(x) = 1 - \Phi(-x)$  and a standard result for Gaussian moments derived in [Rasmussen and Williams, 2006, p. 74]. This expected value equals the point estimate of the likelihood function given in [Gutmann and Corander, 2016].

A formula for the variance can be obtained similarly and the result is

$$\begin{aligned} \mathbb{V}(p(\theta^*)) &= \mathbb{E}(p(\theta^*)^2) - (\mathbb{E}(p(\theta^*)))^2 \\ &= \int_{-\infty}^{\infty} \Phi^2\left(\frac{\varepsilon - f}{\sigma_n}\right) \mathcal{N}(f | m(\theta^*), v^2(\theta^*)) df \\ &\quad - (\mathbb{E}(p(\theta^*)))^2. \end{aligned} \quad (9)$$

There appears to be no closed form solution and thus numerical integration is needed. However, we can use a transformation of variables  $g = (f - m(\theta^*))/v(\theta^*)$  to reformulate the equation. Additionally, the fact that

$$\begin{aligned} \int_{-\infty}^{\infty} \Phi^2(a + bx) \mathcal{N}(x | 0, 1) dx &= \Phi\left(\frac{a}{\sqrt{1 + b^2}}\right) \\ &\quad - 2T\left(\frac{a}{\sqrt{1 + b^2}}, \frac{1}{\sqrt{1 + 2b^2}}\right), \quad a, b \in \mathbb{R}, \end{aligned} \quad (10)$$

where  $T(\cdot, \cdot)$  is Owen's t-function which satisfies

$$T(h, a) = \frac{1}{2\pi} \int_0^a \frac{e^{-h^2(1+x^2)/2}}{1+x^2} dx, \quad (11)$$

<sup>3</sup>This notation should not be confused with probability distribution function which is always denoted with  $\pi(\cdot)$  in this paper. From now on, we will also ignore the conditioning on the data to simplify notation.

(see [Owen, 1956, 1980]) and some straightforward manipulations lead to the following formula for the variance of  $p(\theta^*)$

$$\begin{aligned} \mathbb{V}(p(\theta^*)) &= \Phi\left(\frac{\varepsilon - m(\theta^*)}{\sqrt{\sigma_n^2 + v^2(\theta^*)}}\right) \Phi\left(\frac{m(\theta^*) - \varepsilon}{\sqrt{\sigma_n^2 + v^2(\theta^*)}}\right) \\ &\quad - \frac{1}{\pi} \int_0^{\frac{\sigma_n}{\sqrt{\sigma_n^2 + 2v^2(\theta^*)}}} \frac{e^{-\frac{1}{2}\left(\frac{\varepsilon - m(\theta^*)}{\sqrt{\sigma_n^2 + v^2(\theta^*)}}\right)^2 (1+x^2)}}{1+x^2} dx. \end{aligned} \quad (12)$$

While this seems complicated, an efficient algorithm to evaluate the Owen's t-function exists [Patefield and Tandy, 2000].

It is of interest to examine when the variance in Equation (12) is large. If a parameter  $\theta^*$  satisfies  $m(\theta^*) = \varepsilon$ , then the first term of Equation (12) is maximised, and in this case the second term is maximised when  $v^2(\theta^*)$  is maximised. On the other hand, if  $m(\theta^*) \gg \varepsilon$  but  $v(\theta^*) \gg |m(\theta^*) - \varepsilon|$ , the first term in (12) is approximately maximised and the second term is also close to its maximum value, especially if also  $v(\theta^*) \gg \sigma_n$ . Because the ABC threshold  $\varepsilon$  is usually chosen very small, we thus conclude that the variance in Equation (12) tends to be high in regions where the mean of the discrepancy  $m(\theta^*)$  is small and/or the variance of the latent function  $v^2(\theta^*)$  is large relative to the mean function.

We can also derive the cumulative distribution function and other statistics for the acceptance probability. The derivations are included as supplementary material. For instance, the cdf is

$$F_{p(\theta^*)}(z) = \Phi\left(\frac{\sigma_n \Phi^{-1}(z) + m(\theta^*) - \varepsilon}{v(\theta^*)}\right), \quad (13)$$

if  $z \in (0, 1)$ , and zero if  $z \leq 0$ , and 1 if  $z \geq 1$ . This formula enables the computation of quantiles which can be used for assessing the uncertainty via credible intervals. Setting  $\alpha = F_{p(\theta^*)}(z)$ , where  $\alpha \in (0, 1)$  and solving for  $z$  yields the  $\alpha$ -quantile that was already used in Figure 1a,

$$z_\alpha = \Phi\left(\frac{v(\theta^*) \Phi^{-1}(\alpha) - m(\theta^*) + \varepsilon}{\sigma_n}\right). \quad (14)$$

From the above equation we see, for example, that the median is given by  $\Phi((\varepsilon - m(\theta^*))/\sigma_n)$ .

### 3.3 UNCERTAINTY IN HYPERPARAMETERS

This far we have assumed that the GP hyperparameters  $\phi$  are fixed and known but usually they are not known. The MAP-estimate can be used in the place of the fixed values in the previous formulae. The MAP-estimate can

computed by maximising the logarithm of the marginal likelihood

$$\hat{\phi} = \arg \max_{\phi} \left( \log \pi(\phi) - \frac{1}{2} \Delta^T (k(\theta, \theta) + \sigma_n^2 \mathbf{I})^{-1} \Delta - \frac{1}{2} \log \det(k(\theta, \theta) + \sigma_n^2 \mathbf{I}) \right), \quad (15)$$

where  $\pi(\phi)$  is the prior density for GP hyperparameters and where the covariance function in  $k(\theta, \theta)$  depends also on  $\phi$ .

However, this approach causes underestimation of the variance of  $p(\theta^*)$  as is illustrated in the Figure 7b of the supplementary material. Taking the uncertainty in the GP hyperparameters into account can be done by Monte Carlo sampling ([Murray and Adams, 2010]) or by central composite design ([Rue et al., 2009, Vanhatalo et al., 2010]), which we use here. Briefly, in central composite design (CCD) certain design points  $\phi^i$  are chosen and each of them is given a weight  $\omega^i \propto \pi(\phi^i | D_t) \gamma^i \propto \pi(D_t | \phi^i) \pi(\phi^i) \gamma^i$ , where  $\gamma^i$  is a design weight. This approach has the advantage that the amount of design points grows only moderately with increased dimension and has been shown to yield good accuracy in practice. Further details on choosing the design points and their weights are given in [Vanhatalo et al., 2010].

Taking into account the uncertainty in  $\phi$  leads to the following calculations. Using the law of total expectation yields

$$\begin{aligned} \mathbb{E}(p(\theta^*)) &= \mathbb{E}_{\phi} \mathbb{E}_{p(\theta^*)} (p(\theta^*) | \phi) \\ &\approx \sum_i \omega^i \mathbb{E}_{p(\theta^*)} (p(\theta^*) | \phi^i) \\ &= \sum_i \omega^i \Phi(a(\theta^*, \phi^i)), \end{aligned} \quad (16)$$

where the grid points and the corresponding weights are  $\phi^i$  and  $\omega^i$ , respectively, and where  $a(\theta^*, \phi^i) = (\varepsilon - m(\theta^* | \phi^i)) / \sqrt{(\sigma_n^2)^i + v^2(\theta^* | \phi^i)}$ . Similarly, for the variance we get

$$\begin{aligned} \mathbb{V}(p(\theta^*)) &= \mathbb{E}(p(\theta^*)^2) - (\mathbb{E}(p(\theta^*)))^2 \\ &= \mathbb{E}_{\phi} \mathbb{E}_{p(\theta^*)} (p(\theta^*)^2 | \phi) - (\mathbb{E}_{\phi} \mathbb{E}_{p(\theta^*)} (p(\theta^*) | \phi))^2 \\ &\approx \sum_i \omega^i (\Phi(a(\theta^*, \phi^i)) - 2T(a(\theta^*, \phi^i), b(\theta^*, \phi^i))) \\ &\quad - \left( \sum_i \omega^i \Phi(a(\theta^*, \phi^i)) \right)^2, \end{aligned} \quad (17)$$

where  $b(\theta^*, \phi^i) = (\sigma_n^i)^2 / \sqrt{(\sigma_n^2)^i + 2v^2(\theta^* | \phi^i)}$ .

### 3.4 EFFICIENT PARAMETER ACQUISITION

Given  $t$  evaluations  $D_t$ , we propose to run the simulator model next with the parameter value that maximises the

variance of the likelihood approximation. That is

$$\theta_{t+1} = \arg \max_{\theta \in \Theta} \mathbb{V}(p_t(\theta)), \quad (18)$$

where  $\mathbb{V}(p_t(\theta))$  is given either by Equation (12) or (17) and subscript  $t$  is used here to emphasise that this quantity is computed using the training data  $D_t$ . We call this new strategy 'an acquisition rule' according to the nomenclature in the Bayesian optimisation literature, although our aim is not to optimise the discrepancy but to minimise our uncertainty in the ABC posterior approximation. The derivatives of the variance for e.g. gradient-based optimisation are available (Supplementary material).

So far we have assumed a uniform prior for the parameters of the simulation model. However, because  $\mathbb{E}(\pi(\theta^*) p_t(\theta^*)) = \pi(\theta^*) \mathbb{E}(p_t(\theta^*))$  and  $\mathbb{V}(\pi(\theta^*) p_t(\theta^*)) = \pi^2(\theta^*) \mathbb{V}(p_t(\theta^*))$ , where the expectation and variance are taken with respect to the random variable  $p_t(\theta^*)$ , using any priors with bounded support is straightforward. Specifically, the acquisition strategy in Equation (18) becomes

$$\theta_{t+1} = \arg \max_{\theta \in \Theta} \pi^2(\theta) \mathbb{V}(p_t(\theta)) \quad (19)$$

$$= \arg \max_{\theta \in \Theta} \left( \log \pi(\theta) + \log \sqrt{\mathbb{V}(p_t(\theta))} \right). \quad (20)$$

Furthermore, if the prior is a proper density so that far enough in the tails the objective function in Equation (19) goes to zero, the requirement of the bounded support can be relaxed, although in practice many (global) optimisation methods do require the specification of a box-domain.

To encourage even further exploration, as in [Gutmann and Corander, 2016], we also consider a stochastic variant of the deterministic acquisition function in Equation (20). Specifically, we generate the evaluation point randomly according to the variance surface. That is, instead of finding the maximiser, we generate  $\theta_{t+1} \sim \pi_t(\theta)$ , where  $\pi_t(\theta) \propto \pi^2(\theta) \mathbb{V}(p_t(\theta))$ . It is easy to see that if the prior is bounded and proper i.e.  $\pi(\theta) < \infty$  and  $\int_{\Theta} \pi(\theta) d\theta = 1$ , then the variance surface defines a valid probability density (up to normalisation). This strategy requires generating random samples from  $\pi_t(\theta)$ , which is usually more challenging than optimisation. We do this by grid sampling in small dimensions and by MCMC in higher dimensions. Sampling or optimising the variance function can be typically done fast compared to the time required to run the simulation model.

The stochastic acquisition rule is reminiscent of Thompson sampling, but it is actually quite different. In our method, acquisitions are randomly chosen from the probability distribution which is proportional to the (point-wise) variance of the approximate posterior density. In



Thompson sampling, instead, one generates a posterior density realisation from the model, and chooses the next point as the maximiser of this density.

## 4 EXPERIMENTS

We compare the proposed acquisition rules to commonly used BO strategies: expected improvement (EI) and lower confidence bound (LCB) criterion, see e.g. [Shahriari et al., 2015]. We use the same trade-off parameter for LCB as [Gutmann and Corander, 2016], but unlike them, we consider the deterministic LCB rule. As a simple baseline, we draw points sequentially from the uniform distribution, abbreviated as "unif". We included also the probability of improvement (PI) strategy in preliminary experiments, but it resulted in very poor estimates and was therefore excluded from the comparisons.

We call our acquisition rule with the MAP estimate for GP hyperparameters  $\phi$  as "maxvar" and the stochastic version as "rand\_maxvar". Similarly, the names "int\_maxvar" and "rand\_int\_maxvar" are used if the uncertainty in  $\phi$  is taken into account by CCD integration. We use MATLAB 2015b and GPstuff 4.6 [Vanhatalo et al., 2013] for GP fitting. Total variation (TV) distance is mostly used for the accuracy of the posterior approximation when the ground-truth is available. In high dimensions the average of TVs of marginal densities is used. The point estimates of the ABC posterior densities for the comparisons are computed using the likelihood in Equations (8) and (16).

### 4.1 EXPERIMENTS ON SYNTHETIC 2D DATA

To compare different acquisition strategies first without the need to actually handle different simulation models, we construct "synthetic" discrepancies by adding Gaussian noise to certain parametric curves, and assume the true discrepancy follows these assumptions. The true posterior is computed using the likelihood function given by Equation (7) with a small predefined threshold  $\varepsilon$ . As test cases we consider 1) a unimodal density with positive correlation, 2) a bimodal density, and 3) a banana shaped density, all with a uniform prior (see Supplement for details). The initial training set size is 10 and the threshold is the 0.01th quantile of the realised discrepancies, updated during the acquisitions. Results for two alternative approaches are shown in Supplementary material.

The results are shown in Figure 2. We see that the proposed strategies, especially rand\_maxvar and rand\_int\_maxvar, are consistently the top-performing methods, matched only by LCB in the 'banana' test case

and by maxvar in the 'unimodal' test case. Of the common alternatives, LCB is clearly the best.

### 4.2 GAUSSIAN SIMULATION MODEL

A simple Gaussian simulator model is used to study the effect of prior strength and the dimension of the parameter space. Data are generated independently from  $\mathbf{x}_i \sim \mathcal{N}(\cdot | \boldsymbol{\theta}, \boldsymbol{\Sigma})$ ,  $i = 1, \dots, n$ , where  $\boldsymbol{\theta} \in \Theta = [0, 8]^p$  needs to be estimated and covariance matrix  $\boldsymbol{\Sigma}$  is known. If  $\boldsymbol{\theta} \sim \mathcal{N}(\mathbf{a}, \mathbf{B})$  (truncated to  $\Theta$ ), the true posterior is  $\mathcal{N}(\boldsymbol{\theta} | \mathbf{a}^*, \mathbf{B}^*)$  (truncated to  $\Theta$ ), where  $\mathbf{a}^* = \mathbf{B}^*(\mathbf{B}^{-1}\mathbf{a} + n\boldsymbol{\Sigma}^{-1}\bar{\mathbf{x}}_{obs})$ ,  $\mathbf{B}^* = (\mathbf{B}^{-1} + n\boldsymbol{\Sigma}^{-1})^{-1}$  and  $\bar{\mathbf{x}}_{obs}$  is the sample mean. As discrepancy, we use the Mahalanobis distance  $\Delta_{\boldsymbol{\theta}} = ((\bar{\mathbf{x}}_{obs} - \bar{\mathbf{x}}_{\boldsymbol{\theta}})^T \boldsymbol{\Sigma}^{-1} (\bar{\mathbf{x}}_{obs} - \bar{\mathbf{x}}_{\boldsymbol{\theta}}))^{1/2}$ .

**Strength of prior:** In the first experiment we set  $p = 2$ ,  $n = 5$ ,  $\boldsymbol{\Sigma}_{ii} = 1$ , and  $\boldsymbol{\Sigma}_{ij} = 0.5$  for  $i \neq j$ . The initial training set size is 10 and the threshold is fixed to 0.1. The true data mean of  $\boldsymbol{\theta}$  is  $[2, 2]^T$ . We use a (truncated) Gaussian prior with mean  $\mathbf{a} = [5, 5]^T$  and covariance  $\mathbf{B} = b^2 \mathbf{I}$ . We vary  $b$ , allowing us to study the impact of prior strength relative to the likelihood. Figure 3 shows the results, and we see that the proposed acquisition rules perform consistently well regardless the strength of prior, and focus the evaluations on the posterior modal region. On the other hand, LCB samples where the discrepancies are small, i.e. in areas of high likelihood, leading to sub-optimal posterior estimation. Figure 3b shows that we also avoid unnecessary evaluations on the boundary. Curiously, the unif rule works well when prior information is strong.

**High-dimensional test cases:** Next we investigate the effect of dimension  $p$ . The settings are as before, except that now we use uniform priors supported on  $\Theta$  and adaptive thresholds described in Section 4.1. Further,  $n = 15$ , and the initial training set sizes are 20 (3d) and 40 (6d and 10d). Adaptive MCMC (with multiple chains) is used to sample from the posterior estimates and, in the case of rand\_maxvar, from the variance surface  $\pi_t(\boldsymbol{\theta}^*)$ . Figure 4 shows the results. With  $p \leq 6$ , the rand\_maxvar is again most accurate. However, in 10d it suffers from instability in MCMC convergence. Detailed examination shows that the method often produces multimodal posterior estimates which makes the sampling difficult. Such densities are likely a result of the random acquisitions. Namely, even if there is large uncertainty within some region, it can happen that no evaluations occur there during the available iterations, due to the curse of dimensionality. Similar issues were rarely observed with other strategies. The maxvar strategy was most accurate in 10d suggesting that in high dimensions the deterministic strategy should be preferred over the stochastic one.

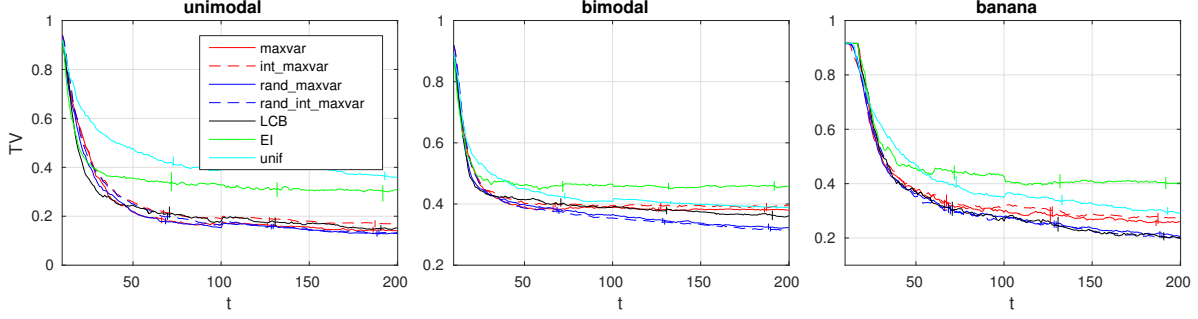


Figure 2: Median of the TV distance between the estimated and true posterior over 100 experiments. Vertical lines show the 95% confidence interval of the median computed using the bootstrap.

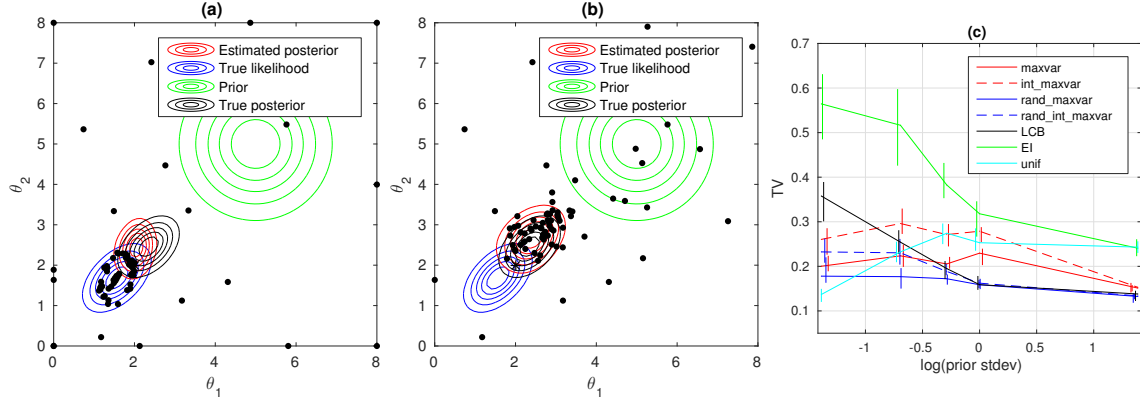


Figure 3: Acquired training data locations (black dots) for (a) LCB, (b) rand\_maxvar after 70 acquisitions. As discussed in [Gutmann and Corander, 2016], the LCB strategy ignores prior information which leads to suboptimal selection of evaluation locations. (c) Median TV between the estimated and true posterior as a function of the standard deviation (stdev) of the Gaussian prior over 100 experiments and after 200 evaluations (small stdev corresponds to strong prior information).

### 4.3 REALISTIC EXAMPLES

We consider the Lotka-Volterra model and a model of bacterial infections in day care centers.

**Lotka-Volterra model:** The Lotka-Volterra (LV) model [Toni et al., 2009] is described by differential equations  $x_1'(t) = \theta_1 x_1(t) - x_1(t)x_2(t)$  and  $x_2'(t) = \theta_2 x_1(t)x_2(t) - x_2(t)$ , where  $x_1(t)$  and  $x_2(t)$  describe the evolution of prey and predator populations as a function of time  $t$ , respectively, and  $\theta$  is the unknown parameter to be estimated. We use a similar experiment design as in [Toni et al., 2009] but with discrepancy  $\Delta_{\theta} = \log \sum_{ij} |x_j^{\text{obs}}(t_i) - x_j^{\text{mod}}(t_i, \theta)|$ , where  $x_j^{\text{obs}}(t_i)$  for  $j \in \{1, 2\}$  denote the noisy observations at times  $t_i$ , and  $x_j^{\text{mod}}(t_i, \theta)$  are the corresponding predictions. In comparisons we use the uniform prior with support on  $[0, 5]^2$ , and the exact posterior distribution as the baseline. The results in Figure 5 show that the rand\_maxvar strategy produces the most accurate posterior approximations followed by LCB and the maxvar strategy.

**Bacterial infections model:** This model describes transmission dynamics of bacterial infections in day care centers. This model has three parameters: an internal infection parameter  $\beta \in [0, 11]$ , an external infection parameter  $\Lambda \in [0, 2]$  and a co-infection parameter  $\theta \in [0, 1]$ . Details of the model and data are described in [Numminen et al., 2013]. The true posterior is not available and thus an approximate posterior computed using PMC-ABC algorithm with over two million simulations is used as the ground-truth [Numminen et al., 2013]. We use the same experiment design and discrepancy as [Gutmann and Corander, 2016], who used the model to illustrate their approach. Specifically, the initial training data size is 20 and uniform prior is used.

Figure 6 shows the results. Unlike other test cases, the additional exploration of stochastic rand\_maxvar causes wider credible intervals and thus poorer posterior approximations than other methods. Similarly, [Gutmann and Corander, 2016] obtained conservative estimates with a stochastic variant of their LCB acquisition rule.

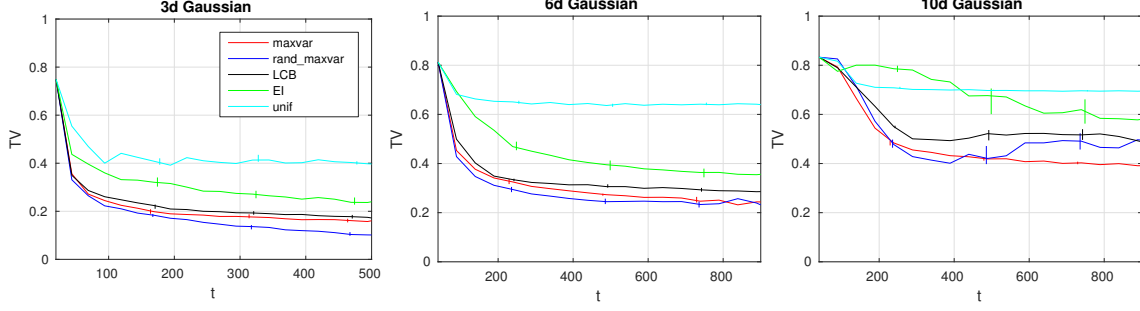


Figure 4: Median of the average marginal TVs between the estimated approximate density and true posterior over 100 experiments in the 3d, 6d and 10d Gaussian toy simulation model.

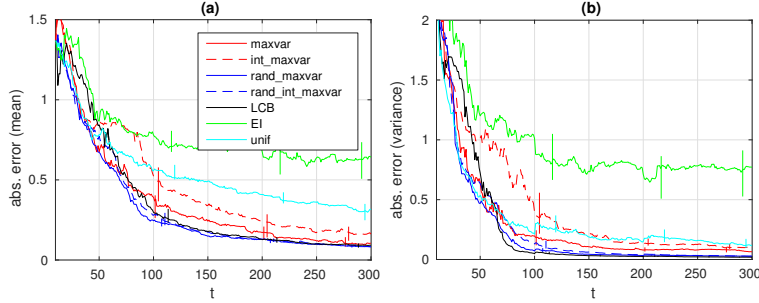


Figure 5: Median of the mean absolute error in the (a) posterior mean, and (b) posterior variance over 100 experiments in the Lotka-Volterra model.

However, with the other acquisition rules we obtain results that are very close to the ground-truth, although using only a few hundred evaluations. Interestingly, the EI strategy also works well in this particular example although it has high variability and occasionally produces too narrow posterior estimates. However, (deterministic) maxvar and (deterministic) LCB produce the most accurate and stable estimates.

## 5 DISCUSSION

The experiments demonstrated that the proposed acquisition rules perform better or, at worse, similarly to alternatives in producing an accurate ABC posterior with few simulations. While not designed for ABC, the LCB criterion also worked surprisingly well. This motivates further study of the connection between the bandit setting and model-based ABC inference. On the other hand, EI (and PI) often performed poorly, and proposed insufficiently many evaluations from the whole modal area. While integrating over the uncertainty in GP hyperparameters improves the accuracy of the variance computation, the resulting acquisition rules (int\_maxvar and rand\_int\_maxvar) did not bring clear improvements over the corresponding acquisition rules with MAP-estimate (maxvar and rand\_maxvar).

One challenge of the proposed acquisition rules is that the threshold must be chosen. We used a heuristic approach and set the threshold to the 0.01th quantile of the realised discrepancies. We also considered other choices but this approach worked well. In principle, the strategy for selecting the threshold could also vary during the iterations. However, while some ABC methods bypass selecting the threshold, they may not be applicable when the budget for simulations is very small.

We used the zero mean GP in our experiments. While [Wilkinson, 2014, Drovandi et al., 2015, Gutmann and Corander, 2016] considered certain parametric mean functions which might help focusing the simulations on the modal area, our choice is a safe option. Namely, if there is a large region containing no simulations, the discrepancy tends to zero there. Thus, the uncertainty will be high in the region, attracting future simulations.

Our approach is based on modelling the discrepancy with a GP. However, this approach may not be optimal if the Gaussian assumption is violated [Gutmann and Corander, 2016, Järvenpää et al., 2016]. Non-Gaussian measurement models can be used but the variance in Equation (12) or (17) may become costly to evaluate. We outline other modelling approaches, also suitable for the framework outlined in Section 2, in the supplementary



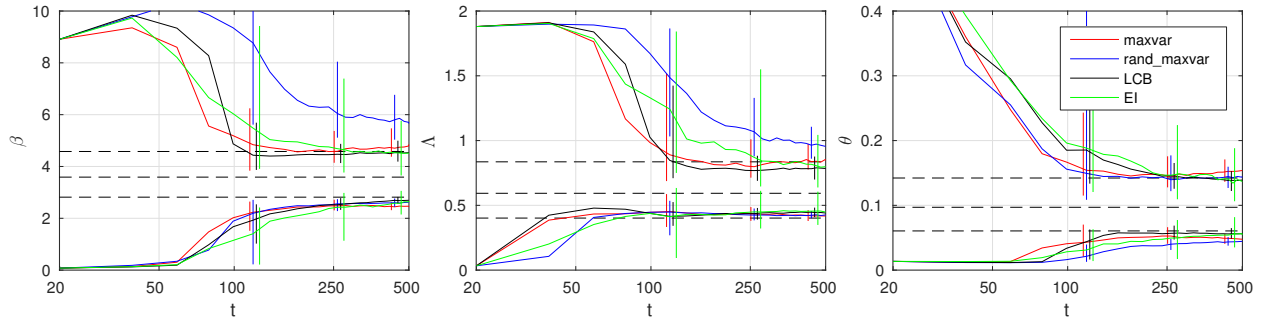


Figure 6: Comparison of the 95% credible interval estimates in the bacterial model. The black dashed lines show the ground truth by [Numminen et al., 2013] and the vertical lines show the 75% interval of the realisations over 100 experiments.

material.

An alternative to the proposed stochastic acquisition rule is to sample new evaluation locations from the current posterior estimate. This approach seems to work well in some scenarios but no systematic comparison was done. However, the posterior estimate could get stuck to a poor region due to an "unlucky" discrepancy realisation, after which new evaluations would be focused on this seemingly good region only.

## 6 CONCLUSIONS

We considered the challenging problem of performing Bayesian inference when the likelihood function cannot be evaluated and simulating data from the statistical model is costly. We proposed to quantify the uncertainty in the approximate posterior due to the limited budget of simulations and to design the simulations so as to minimise the expected uncertainty in the approximation to the posterior. Two tractable approximate strategies were developed under GP modelling assumptions. Experiments demonstrated the advantages of the proposed approach, and pointed to several directions for future work. Other surrogate models and principled approaches for selecting the threshold could be investigated. Also, developing batch acquisition strategies that attempt to minimise the uncertainty in the ABC posterior would allow parallelised inference for computationally very costly simulation models.

## Acknowledgements

This work was funded by the Academy of Finland (grants no. 286607 and 294015 to PM). We acknowledge the computational resources provided by Aalto Science-IT project.

## References

- M. A. Beaumont, J.-M. Cornuet, J.-M. Marin, and C. P. Robert. Adaptive approximate Bayesian computation. *Biometrika*, 96(4):983–990, 2009.
- H. Bijl, T. B. Schön, J. van Wingerden, and M. Verhaegen. A sequential Monte Carlo approach to Thompson sampling for Bayesian optimization. *arXiv:1604.00169*, 2016.
- M. G. B. Blum, M. A. Nunes, D. Prangle, and S. A. Sisson. A comparative review of dimension reduction methods in approximate Bayesian computation. *Statistical Science*, 28(2):189–208, 2013.
- P. J. Diggle and R. J. Gratton. Monte Carlo Methods of Inference for Implicit Statistical Models. *Journal of the Royal Statistical Society. Series B (Methodological)*, 46(2):193–227, 1984.
- C. C. Drovandi, M. T. Moores, and R. J. Boys. Accelerating pseudo-marginal MCMC using Gaussian processes. Available at <http://eprints.qut.edu.au/90973/>. Accessed 13-3-2017, 2015.
- Y. Fan, D. J. Nott, and S. A. Sisson. Approximate Bayesian computation via regression density estimation. *Stat*, 2(1):34–48, 2013.
- P. Fearnhead and D. Prangle. Constructing summary statistics for approximate Bayesian computation: Semi-automatic approximate Bayesian computation. *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, 74(3):419–474, 2012.
- M. U. Gutmann and J. Corander. Bayesian optimization for likelihood-free inference of simulator-based statistical models. *Journal of Machine Learning Research*, 17(125):1–47, 2016.
- P. Hennig and C. J. Schuler. Entropy Search for Information-Efficient Global Optimization. *Journal of Machine Learning Research*, 13(1999):1809–1837, 2012.

- J. M. Hernández-Lobato, M. W. Hoffman, and Z. Ghahramani. Predictive Entropy Search for Efficient Global Optimization of Black-box Functions. *Advances in Neural Information Processing Systems* 28, pages 1–9, 2014.
- F. Jabot, G. Lagarrigues, B. Courbaud, and N. Dumoulin. A comparison of emulation methods for Approximate Bayesian Computation. *arXiv:1412.7560*, 2014.
- M. Järvenpää, M. Gutmann, A. Vehtari, and P. Marttinen. Gaussian process modeling in approximate Bayesian computation to estimate horizontal gene transfer in bacteria. *arXiv:1610.06462*, 2016.
- K. Kandasamy, J. Schneider, and B. Póczos. Bayesian active learning for posterior estimation. In *International Joint Conference on Artificial Intelligence*, pages 3605–3611, 2015.
- M. Lenormand, F. Jabot, and G. Deffuant. Adaptive approximate Bayesian computation for complex models. *Computational Statistics*, 28(6):2777–2796, 2013.
- J. Lintusaari, M. U. Gutmann, R. Dutta, S. Kaski, and J. Corander. Fundamentals and Recent Developments in Approximate Bayesian Computation. *Systematic biology*, 66(1):e66–e82, 2016.
- J. M. Marin, P. Pudlo, C. P. Robert, and R. J. Ryder. Approximate Bayesian computational methods. *Statistics and Computing*, 22(6):1167–1180, 2012.
- E. Meeds and M. Welling. GPS-ABC: Gaussian Process Surrogate Approximate Bayesian Computation. In *Proceedings of the 30th Conference on Uncertainty in Artificial Intelligence*, 2014.
- I. Murray and R. P. Adams. Slice sampling covariance hyperparameters of latent Gaussian models. *Advances in Neural Information Processing Systems*, 2(1):9, 2010.
- E. Numminen, L. Cheng, M. Gyllenberg, and J. Corander. Estimating the transmission dynamics of streptococcus pneumoniae from strain prevalence data. *Biometrics*, 69(3):748–757, 2013.
- D. B. Owen. Tables for computing bivariate normal probabilities. *The Annals of Mathematical Statistics*, 27(4):1075–1090, 12 1956.
- D. B. Owen. A table of normal integrals. *Communications in Statistics - Simulation and Computation*, 9(4):389–419, 1980.
- G. Papamakarios and I. Murray. Fast e-free inference of simulation models with Bayesian conditional density estimation. In *Advances in Neural Information Processing Systems* 29, 2016.
- M. Patefield and D. Tandy. Fast and accurate Calculation of Owen’s T-Function. *Journal of Statistical Software*, 5(5):1–25, 2000.
- L. F. Price, C. C. Drovandi, A. Lee, and D. J. Nott. Bayesian synthetic likelihood. Available at <http://eprints.qut.edu.au/92795/> Accessed 13-3-2017, 2016.
- C. E. Rasmussen. Gaussian Processes to Speed up Hybrid Monte Carlo for Expensive Bayesian Integrals. *Bayesian Statistics 7*, pages 651–659, 2003.
- C. E. Rasmussen and C. K. I. Williams. *Gaussian Processes for Machine Learning*. The MIT Press, 2006.
- H. Rue, S. Martino, and N. Chopin. Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, 71(2):319–392, 2009.
- B. Shahriari, K. Swersky, Z. Wang, R. P. Adams, and N. de Freitas. Taking the human out of the loop: A review of Bayesian optimization. *Proceedings of the IEEE*, 104(1), 2015.
- T. Toni, D. Welch, N. Strelkowa, A. Ipsen, and M. P. H. Stumpf. Approximate Bayesian computation scheme for parameter inference and model selection in dynamical systems. *Journal of the Royal Society, Interface*, 6(31):187–202, 2009.
- B. M. Turner and T. Van Zandt. A tutorial on approximate Bayesian computation. *Journal of Mathematical Psychology*, 56(2):69–85, 2012.
- J. Vanhatalo, V. Pietiläinen, and A. Vehtari. Approximate inference for disease mapping with sparse Gaussian processes. *Statistics in Medicine*, 29(15):1580–1607, 2010.
- J. Vanhatalo, J. Riihimäki, J. Hartikainen, P. Jylänki, V. Tolvanen, and A. Vehtari. GPstuff: Bayesian modeling with Gaussian processes. *Journal of Machine Learning Research*, 14:1175–1179, 2013.
- Z. Wang, B. Zhou, and S. Jegelka. Optimization as Estimation with Gaussian Processes in Bandit Settings. In *In proceedings of the 19th International Conference on Artificial Intelligence and Statistics*, 2016.
- R. D. Wilkinson. Approximate Bayesian computation (ABC) gives exact results under the assumption of model error. *Statistical Applications in Genetics and Molecular Biology*, 12(2):129–141, 2013.
- R. D. Wilkinson. Accelerating ABC methods using Gaussian processes. In *Proceedings of the 17th International Conference on Artificial Intelligence and Statistics*, 2014.
- S. N. Wood. Statistical inference for noisy nonlinear ecological dynamic systems. *Nature*, 466:1102–1104, 2010.

## A SUPPLEMENTARY MATERIAL

We present some additional derivations and results not shown in the main text. Specifically, Section A.1 contains additional derivations and facts related to the model-based estimation of the approximate likelihood function. Derivatives for the proposed acquisition function and for the posterior approximation are derived in Section A.2. In Section A.3 we present further illustrations and additional results of the experiments. Finally, some alternative approaches and extensions to our methodology are briefly discussed in Section A.4.

### A.1 ADDITIONAL DERIVATIONS

We start by deriving the cdf for the random variable  $p(\theta^*)$  when the uncertainty in the GP hyperparameters  $\phi$  is taken into account. The cdf of  $p(\theta^*)$  evaluated at  $z \in (0, 1)$  is

$$\begin{aligned} F_{p(\theta^*)}(z) &= \mathbb{P}(p(\theta^*) \leq z) = \int \mathbb{P}(p(\theta^*) \leq z | \phi) \pi_\phi(\phi) d\phi = \int \mathbb{P}\left(\Phi\left(\frac{\varepsilon - f(\theta^*)}{\sigma_n}\right) \leq z \mid \phi\right) \pi_\phi(\phi) d\phi \\ &= \int \mathbb{P}\left(f(\theta^*) \geq \varepsilon - \sigma_n \Phi^{-1}(z) \mid \phi\right) \pi_\phi(\phi) d\phi = \int \Phi\left(\frac{\sigma_n \Phi^{-1}(z) + m(\theta^* | \phi) - \varepsilon}{v(\theta^* | \phi)}\right) \pi_\phi(\phi) d\phi, \end{aligned} \quad (21)$$

it is zero if  $z \leq 0$ , and 1 if  $z \geq 1$ . In above, the density  $\pi_\phi(\phi)$  describes our knowledge about the GP hyperparameters  $\phi$  given the training data  $D_t$  (conditioning on data is again ignored to simplify notation). The integral in the equations above is taken over the domain of the GP hyperparameters  $\phi$ . The integral can be approximated using e.g. CCD as discussed in the main text. If the hyperparameters are fixed and  $\pi_\phi(\phi)$  is replaced with a point mass, one obtains the Equation (13).

A formula for the pdf can be obtained by differentiating the cdf. However, first we realise that  $z = \Phi(\Phi^{-1}(z)) \implies 1 = \frac{dz}{dz} = \frac{d}{dz} \Phi(\Phi^{-1}(z)) = \Phi'(\Phi^{-1}(z))(\Phi^{-1})'(z)$  for  $z \in (0, 1)$ . This fact further leads to

$$(\Phi^{-1})'(z) = \frac{1}{\Phi'(\Phi^{-1}(z))} = \frac{1}{\mathcal{N}(\Phi^{-1}(z) | 0, 1)} = \sqrt{2\pi} e^{(\Phi^{-1}(z))^2/2}. \quad (22)$$

Using the formula (22) allows to compute

$$\begin{aligned} \pi_{p(\theta^*) | \phi}(z) &= \frac{\partial}{\partial z} F_{p(\theta^*) | \phi}(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{(\sigma_n \Phi^{-1}(z) + m(\theta^* | \phi) - \varepsilon)^2}{2v^2(\theta^* | \phi)}} \frac{\partial}{\partial z} \frac{\sigma_n \Phi^{-1}(z) + m(\theta^* | \phi) - \varepsilon}{v(\theta^* | \phi)} \\ &= \frac{\sigma_n}{v(\theta^* | \phi)} e^{\frac{(\Phi^{-1}(z))^2}{2} - \frac{(\sigma_n \Phi^{-1}(z) + m(\theta^* | \phi) - \varepsilon)^2}{2v^2(\theta^* | \phi)}} \\ &= \begin{cases} \frac{\sigma_n}{v(\theta^* | \phi)} e^{\frac{(\varepsilon - m(\theta^* | \phi))^2}{2(\sigma_n^2 - v^2(\theta^* | \phi))}} e^{-\frac{\sigma_n^2 - v^2(\theta^* | \phi)}{2v^2(\theta^* | \phi)} \left(\Phi^{-1}(z) - \frac{(\varepsilon - m(\theta^* | \phi))\sigma_n}{\sigma_n^2 - v^2(\theta^* | \phi)}\right)^2}, & \text{if } \sigma_n \neq v(\theta^* | \phi), \\ e^{-\frac{(\varepsilon - m(\theta^* | \phi))^2}{2v^2(\theta^* | \phi)}} e^{\frac{\varepsilon - m(\theta^* | \phi)}{v(\theta^* | \phi)} \Phi^{-1}(z)}, & \text{if } \sigma_n = v(\theta^* | \phi), \end{cases} \end{aligned} \quad (23)$$

for  $z \in (0, 1)$  and it is zero elsewhere. Finally, the pdf is obtained by marginalising the GP hyperparameters, that is

$$\pi_{p(\theta^*)}(z) = \int \pi_{p(\theta^*) | \phi}(z) \pi_\phi(\phi) d\phi. \quad (24)$$

The integral in the Equation (24) can be approximated as before.

The derivations of the mean and variance for  $p(\theta^*)$  were already outlined in Section 3.2. The joint pdf and covariance between any two evaluation points can be also derived but we do not present these formulae here. The quantiles can be computed as in Equation (14). If the uncertainty in the GP hyperparameters is taken into account, then numerical root finding such as bisection search is required for computing the quantiles. If the GP hyperparameters are fixed, the differential entropy at  $\theta^*$  is given by

$$H(p(\theta^*)) = \log \frac{v(\theta^*)}{\sigma_n} - \frac{(\varepsilon - m(\theta^*))^2 + v(\theta^*)^2 - \sigma_n^2}{2\sigma_n^2}. \quad (25)$$

Instead of evaluating where the variance of  $p(\boldsymbol{\theta}^*)$  is highest, one could evaluate where the differential entropy is highest. However, we do not investigate this alternative approach here. We also omit the derivation because this formula follows straightforwardly by using Equation (23) and the definition of differential entropy although the derivation is somewhat tedious. If the uncertainty in the GP hyperparameters is taken into account, then the differential entropy can be computed numerically using the Equation (24) and the definition of differential entropy.

Inspecting the formula (23) shows that if  $\sigma_n > v(\boldsymbol{\theta}^* | \phi)$ , then the mode of  $p(\boldsymbol{\theta}^* | \phi)$  is at  $z = \Phi((\varepsilon - m(\boldsymbol{\theta}^* | \phi)) / (\sigma_n - v^2(\boldsymbol{\theta}^* | \phi) / \sigma_n))$ . Unsurprisingly, if  $m(\boldsymbol{\theta}^* | \phi)$  is large enough, then there is a mode near  $z = 0$ . However, if  $\sigma_n = v(\boldsymbol{\theta}^* | \phi)$  and  $m(\boldsymbol{\theta}^* | \phi) > \varepsilon$ , then the pdf goes to infinity as  $z \rightarrow 0^+$ . Interestingly, if  $\sigma_n < v(\boldsymbol{\theta}^* | \phi)$ , then the pdf goes to infinity both as  $z \rightarrow 0^+$  and  $z \rightarrow 1^-$ .

## A.2 DERIVATIVES OF THE ACQUISITION FUNCTION AND POSTERIOR APPROXIMATION

We compute the derivatives of the proposed acquisition function with respect to the unknown parameter vector  $\boldsymbol{\theta}^*$ . We take into account the uncertainty in the GP hyperparameters but if some point estimate is used instead, the formulae can be simplified by ignoring the summations, setting  $\omega^i = \omega^1 = 1$  and replacing  $\phi^i$  with the point estimate. First we denote

$$I_1(\boldsymbol{\theta}^*) = \sum_i \omega^i \Phi(a(\boldsymbol{\theta}^*, \phi^i)) - \left( \sum_i \omega^i \Phi^2(a(\boldsymbol{\theta}^*, \phi^i)) \right)^2, \quad (26)$$

$$I_2(\boldsymbol{\theta}^*) = \frac{1}{\pi} \sum_i \omega^i \int_0^{b(\boldsymbol{\theta}^*, \phi^i)} \frac{e^{-\frac{1}{2}a^2(\boldsymbol{\theta}^*, \phi^i)(1+x^2)}}{1+x^2} dx, \quad (27)$$

where

$$a(\boldsymbol{\theta}^*, \phi^i) = \frac{\varepsilon - m(\boldsymbol{\theta}^* | \phi^i)}{\sqrt{(\sigma_n^2)^i + v^2(\boldsymbol{\theta}^* | \phi^i)}}, \quad b(\boldsymbol{\theta}^*, \phi^i) = \frac{(\sigma_n)^i}{\sqrt{(\sigma_n^2)^i + 2v^2(\boldsymbol{\theta}^* | \phi^i)}}. \quad (28)$$

The Equation (17) shows that

$$\pi^2(\boldsymbol{\theta}^*) \mathbb{V}(p(\boldsymbol{\theta}^*)) \approx \pi^2(\boldsymbol{\theta}^*) (I_1(\boldsymbol{\theta}^*) - I_2(\boldsymbol{\theta}^*)). \quad (29)$$

Differentiating the Equation (29) with respect to the parameter vector yields

$$\frac{\partial}{\partial \boldsymbol{\theta}^*} \pi^2(\boldsymbol{\theta}^*) \mathbb{V}(p(\boldsymbol{\theta}^*)) \approx 2\pi(\boldsymbol{\theta}^*) (I_1(\boldsymbol{\theta}^*) - I_2(\boldsymbol{\theta}^*)) \frac{\partial \pi(\boldsymbol{\theta}^*)}{\partial \boldsymbol{\theta}^*} + \pi^2(\boldsymbol{\theta}^*) \left( \frac{\partial I_1(\boldsymbol{\theta}^*)}{\partial \boldsymbol{\theta}^*} - \frac{\partial I_2(\boldsymbol{\theta}^*)}{\partial \boldsymbol{\theta}^*} \right). \quad (30)$$

Computing the derivatives of  $I_1$  produces

$$\begin{aligned} \frac{\partial I_1(\boldsymbol{\theta}^*)}{\partial \boldsymbol{\theta}^*} &= \sum_i \omega^i \frac{\partial}{\partial \boldsymbol{\theta}^*} \Phi(a(\boldsymbol{\theta}^*, \phi^i)) - 2 \left( \sum_i \omega^i \Phi(a(\boldsymbol{\theta}^*, \phi^i)) \right) \left( \sum_i \omega^i \frac{\partial}{\partial \boldsymbol{\theta}^*} \Phi(a(\boldsymbol{\theta}^*, \phi^i)) \right) \\ &= \left( 1 - 2 \sum_i \omega^i \Phi(a(\boldsymbol{\theta}^*, \phi^i)) \right) \sum_i \omega^i \frac{\partial}{\partial \boldsymbol{\theta}^*} \Phi(a(\boldsymbol{\theta}^*, \phi^i)) \\ &= \left( 1 - 2 \sum_i \omega^i \Phi(a(\boldsymbol{\theta}^*, \phi^i)) \right) \sum_i \frac{\omega^i e^{-\frac{1}{2}a^2(\boldsymbol{\theta}^*, \phi^i)}}{\sqrt{2\pi}} \frac{\partial a(\boldsymbol{\theta}^*, \phi^i)}{\partial \boldsymbol{\theta}^*}. \end{aligned} \quad (31)$$

Using the Leibniz integration rule, the derivative of  $I_2$  can be written as

$$\frac{\partial I_2(\boldsymbol{\theta}^*)}{\partial \boldsymbol{\theta}^*} = \frac{1}{\pi} \sum_i \omega^i \left( \frac{e^{-\frac{1}{2}a^2(\boldsymbol{\theta}^*, \phi^i)(1+b^2(\boldsymbol{\theta}^*, \phi^i))}}{1+b^2(\boldsymbol{\theta}^*, \phi^i)} \frac{\partial b(\boldsymbol{\theta}^*, \phi^i)}{\partial \boldsymbol{\theta}^*} + \int_0^{b(\boldsymbol{\theta}^*, \phi^i)} \frac{1}{1+x^2} \frac{\partial}{\partial \boldsymbol{\theta}^*} e^{-\frac{1}{2}a^2(\boldsymbol{\theta}^*, \phi^i)(1+x^2)} dx \right), \quad (32)$$

where the integration is applied elementwise. The second term in the Equation (32) can be further simplified as

$$\begin{aligned}
& \int_0^{b(\boldsymbol{\theta}^*, \phi^i)} \frac{1}{1+x^2} \frac{\partial}{\partial \boldsymbol{\theta}^*} e^{-\frac{1}{2}a^2(\boldsymbol{\theta}^*, \phi^i)(1+x^2)} dx \\
&= - \int_0^{b(\boldsymbol{\theta}^*, \phi^i)} a(\boldsymbol{\theta}^*, \phi^i) \frac{\partial a(\boldsymbol{\theta}^*, \phi^i)}{\partial \boldsymbol{\theta}^*} e^{-\frac{1}{2}a^2(\boldsymbol{\theta}^*, \phi^i)(1+x^2)} dx \\
&= -a(\boldsymbol{\theta}^*, \phi^i) \frac{\partial a(\boldsymbol{\theta}^*, \phi^i)}{\partial \boldsymbol{\theta}^*} e^{-\frac{1}{2}a^2(\boldsymbol{\theta}^*, \phi^i)} \int_0^{b(\boldsymbol{\theta}^*, \phi^i)} e^{-\frac{1}{2}a^2(\boldsymbol{\theta}^*, \phi^i)x^2} dx \\
&= -\sqrt{2\pi} \frac{\partial a(\boldsymbol{\theta}^*, \phi^i)}{\partial \boldsymbol{\theta}^*} e^{-\frac{1}{2}a^2(\boldsymbol{\theta}^*, \phi^i)} (\Phi(a(\boldsymbol{\theta}^*, \phi^i)b(\boldsymbol{\theta}^*, \phi^i)) - \Phi(0)) \\
&= \sqrt{\pi/2} e^{-\frac{1}{2}a^2(\boldsymbol{\theta}^*, \phi^i)} (1 - 2\Phi(a(\boldsymbol{\theta}^*, \phi^i)b(\boldsymbol{\theta}^*, \phi^i))) \frac{\partial a(\boldsymbol{\theta}^*, \phi^i)}{\partial \boldsymbol{\theta}^*}, \tag{33}
\end{aligned}$$

where on the third line we have recognised the integrand as an unnormalised Gaussian pdf.

Finally, straightforward calculations show that

$$\frac{\partial a(\boldsymbol{\theta}^*, \phi^i)}{\partial \boldsymbol{\theta}^*} = -\frac{1}{\sqrt{(\sigma_n^2)^i + v^2(\boldsymbol{\theta}^*, \phi^i)}} \frac{\partial m(\boldsymbol{\theta}^*, \phi^i)}{\partial \boldsymbol{\theta}^*} - \frac{\varepsilon - m(\boldsymbol{\theta}^*, \phi^i)}{2((\sigma_n^2)^i + v^2(\boldsymbol{\theta}^*, \phi^i))^{3/2}} \frac{\partial v^2(\boldsymbol{\theta}^*, \phi^i)}{\partial \boldsymbol{\theta}^*}, \tag{34}$$

$$\frac{\partial b(\boldsymbol{\theta}^*, \phi^i)}{\partial \boldsymbol{\theta}^*} = -\frac{(\sigma_n)^i}{((\sigma_n^2)^i + 2v^2(\boldsymbol{\theta}^*, \phi^i))^{3/2}} \frac{\partial v^2(\boldsymbol{\theta}^*, \phi^i)}{\partial \boldsymbol{\theta}^*}. \tag{35}$$

Derivatives with respect to the GP mean and variance functions  $m$  and  $v^2$  depend on the chosen covariance function and are not shown here.

Derivatives of the posterior approximation with respect to parameter  $\boldsymbol{\theta}$  are useful if e.g. Hamiltonian Monte Carlo is used to sample from this density. The approximate unnormalised posterior is

$$\tilde{\pi}_{\text{ABC}}(\boldsymbol{\theta}^* | x_{\text{obs}}) = \pi(\boldsymbol{\theta}^*) \Phi\left(\frac{\varepsilon - m(\boldsymbol{\theta}^*)}{\sqrt{\sigma_n^2 + v^2(\boldsymbol{\theta}^*)}}\right) = \pi(\boldsymbol{\theta}^*) \Phi(a(\boldsymbol{\theta}^*)), \tag{36}$$

where, as earlier,  $\pi(\boldsymbol{\theta})$  denotes the prior density and  $a$  is defined as in the Equation (28). Differentiating the Equation (36) yields

$$\begin{aligned}
\frac{\partial}{\partial \boldsymbol{\theta}^*} \tilde{\pi}_{\text{ABC}}(\boldsymbol{\theta}^* | x_{\text{obs}}) &= \frac{\partial \pi(\boldsymbol{\theta}^*)}{\partial \boldsymbol{\theta}^*} \Phi(a(\boldsymbol{\theta}^*)) + \pi(\boldsymbol{\theta}^*) \frac{\partial}{\partial \boldsymbol{\theta}^*} \Phi(a(\boldsymbol{\theta}^*)) \\
&= \Phi(a(\boldsymbol{\theta}^*)) \frac{\partial \pi(\boldsymbol{\theta}^*)}{\partial \boldsymbol{\theta}^*} + \frac{\pi(\boldsymbol{\theta}^*) e^{-\frac{1}{2}a^2(\boldsymbol{\theta}^*)}}{\sqrt{2\pi}} \frac{\partial a(\boldsymbol{\theta}^*)}{\partial \boldsymbol{\theta}^*}, \tag{37}
\end{aligned}$$

where the derivative of  $a$  can be computed as in Equation (34). If the uncertainty in GP hyperparameters is taken into account, then the derivation goes similarly.

### A.3 ADDITIONAL DETAILS AND EXPERIMENTS

An additional comparison of the acquisition functions in a one-dimensional toy problem is shown in Figure 7. A simulation model has been run eight times and the acquisition functions for selecting the ninth evaluation location are plotted for comparison. The maxvar acquisition function is plotted for three values of the threshold  $\varepsilon$ . Figure 7 shows that using the MAP-estimate for the GP hyperparameter causes underestimation of the variance. Large threshold, on the other hand, causes the next evaluation to be made in the tail area.

The synthetic test problems in Section 4.1 are designed in the following way. In the "unimodal" example, the mean of the discrepancy is  $m(\boldsymbol{\theta}) = 3\sigma + \boldsymbol{\theta}^T \mathbf{S}\boldsymbol{\theta}$ , where  $\sigma$  is the standard deviation of the additive Gaussian noise,  $\mathbf{S}_{11} = \mathbf{S}_{22} = 1$  and  $\mathbf{S}_{12} = \mathbf{S}_{21} = 0.5$ . In the "bimodal" example we use  $m(\boldsymbol{\theta}) = 3\sigma + \min\{(\boldsymbol{\theta} + \mathbf{1})^T \mathbf{U}(\boldsymbol{\theta} + \mathbf{1}), (\boldsymbol{\theta} - 1.5\mathbf{1})^T \mathbf{V}(\boldsymbol{\theta} - \mathbf{1})\}$ , where  $\mathbf{1} = [1, 1]^T$ ,  $\mathbf{U}_{11} = \mathbf{U}_{22} = 1$ ,  $\mathbf{U}_{12} = \mathbf{U}_{21} = -0.5$ ,  $\mathbf{V}_{11} = 1$ ,  $\mathbf{V}_{22} = 1.5$  and  $\mathbf{V}_{12} = \mathbf{V}_{21} = -0.5$ . The



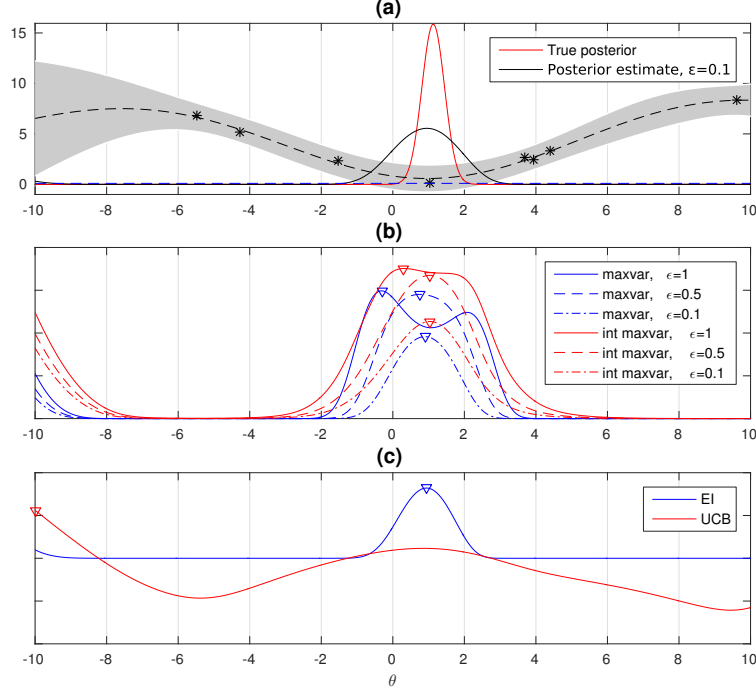


Figure 7: (a) The discrepancy observations (black stars) and the estimate of the posterior density based on the eight training data points (with  $\epsilon = 0.1$ ) as compared to the true posterior. (b) The variance curve is computed using the MAP estimate (maxvar) or CCD integration (int\_maxvar) for three values of the threshold  $\epsilon$ . The scales of the acquisition functions computed with different thresholds are not comparable. (c) EI and UCB criteria (scaled to fit the same figure).

"banana" example is produced using  $m(\theta) = 3\sigma + (1 - \theta_1)^2 + 10(\theta_2 - \theta_1^2)^2$ . The discrepancy is assumed to follow the Gaussian density, that is,  $\Delta\theta \sim \mathcal{N}(m(\theta), \sigma_n^2)$ . The resulting probability densities with  $\sigma = 2$  are illustrated in Figure 8.

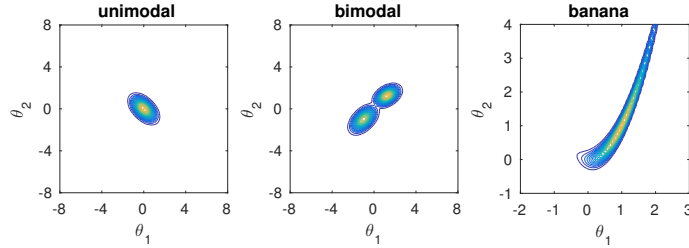


Figure 8: True posterior densities for the synthetic 2d test problems in Section 4.1.

We present additional results for the 2d experiments in Section 4.1. The settings are the same except that the threshold is fixed to  $\epsilon = 0$  so that the differences of the approximation quality between the acquisition methods is solely due to the selection of the evaluation locations. The results are shown in Figure 9. Figure 10 shows the corresponding results when the threshold is determined using the 0.05th quantile.

#### A.4 EXTENSIONS AND ALTERNATIVES

If the likelihood can be evaluated exactly but with high computational cost, modelling the log-likelihood function with a GP may be advantageous. The log-likelihood can also be approximated as in Equation (3), or using parametric assumptions, e.g. the synthetic likelihood [Wood, 2010, Price et al., 2016]. This requires running a batch of simulations with each parameter  $\theta$  considered. The training data  $D_t$  then consists of likelihood evaluations and corresponding

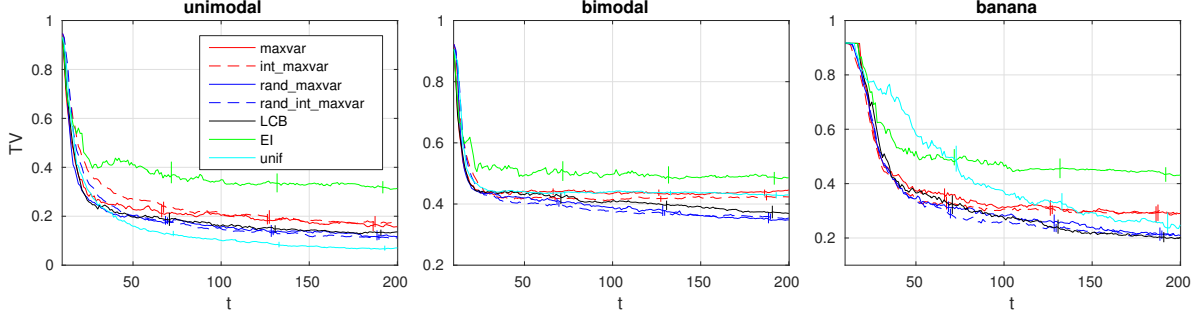


Figure 9: Median of the TV distance between the estimated and the true posterior over 100 experiments. These experiments are as in Figure 2 except that the threshold is fixed. The results are similar as in Figure 2. Interestingly, the uniform strategy produces the best estimates in the case of the unimodal example. The acquisitions are not focused on the modal region but because the modelling assumptions hold everywhere and the parameter space is rather small, the extrapolation works well in this particular case.

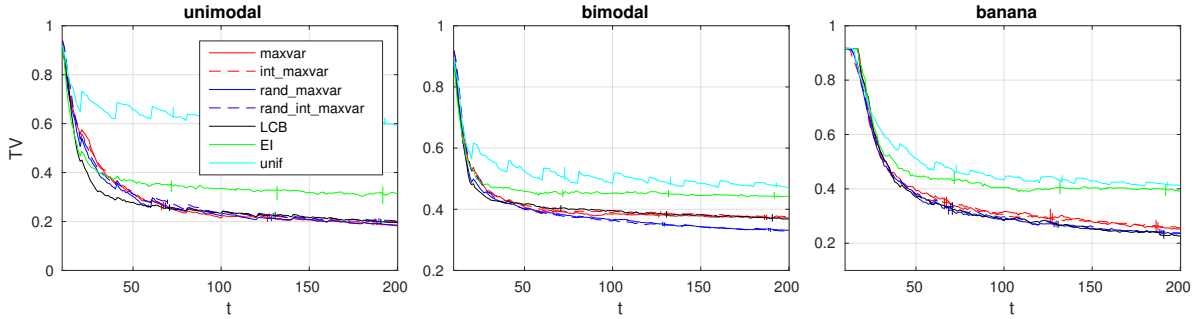


Figure 10: Median of the TV distance between the estimated and the true posterior over 100 experiments. These experiments are as in Figure 2 except that the 0.05th quantile is used for updating the threshold. Larger threshold as in Figure 2 generally produces slightly worse posterior estimates.

parameters  $\theta_i$ . Using the connection between the Normal and log-Normal distributions leads to

$$\mathbb{V}(\pi(\mathbf{x}_{obs} | \theta^*)) \approx \sum_i \omega^i e^{2m(\theta^*, \phi^i) + 2v^2(\theta^*, \phi^i)} - \left( \sum_i \omega^i e^{m(\theta^*, \phi^i) + \frac{1}{2}v^2(\theta^*, \phi^i)} \right)^2, \quad (38)$$

where  $m(\theta^*, \phi^i)$  and  $v^2(\theta^*, \phi^i)$  are as in Equations (5) and (6), but with the observed discrepancies  $\Delta$  replaced by log-likelihood evaluations. Prior information could be taken into account by multiplying Equation (38) with  $\pi^2(\theta^*)$  and new evaluation locations are chosen by maximising the resulting formula. If a point estimate for  $\phi$  is used, then Equation (38) simplifies to  $e^{2m(\theta^*) + v^2(\theta^*)}(e^{v^2(\theta^*)} - 1)$  as in [Kandasamy et al., 2015].

The summary statistics could be assumed independent given the parameters, and modelled separately [Jabot et al., 2014, Meeds and Welling, 2014]. For example, each of some  $m$  summary statistics  $s_i(\theta)$  could be modelled as a function of (some subset of) the parameters  $\theta$  with a GP. The acceptance probability is then obtained from the  $m$  GPs as  $p(\theta^*) = \prod_{i=1}^m \mathbb{P}(|s_i(\theta^*) - s_{obs_i}| \leq \varepsilon_i)$ , where each summary  $s_i$  now requires its own  $\varepsilon$ -parameter. For independent random variables  $x_i, i = 1, \dots, m$ , it holds that  $\mathbb{V}(\prod_{i=1}^m x_i) = \prod_{i=1}^m \mathbb{E}(x_i^2) - \prod_{i=1}^m \mathbb{E}(x_i)^2$ , which allows computing the variance in this case. Further details are not considered here.

Instead of using the uniform (i.e. "0-1") threshold  $\pi_\varepsilon(\mathbf{x}_{obs} | \mathbf{x}) \propto \mathbf{1}_{\Delta(\mathbf{x}_{obs}, \mathbf{x}) \leq \varepsilon}$ , other choices are possible. For instance, one could consider "Gaussian threshold"  $\pi_\varepsilon(\mathbf{x}_{obs} | \mathbf{x}) \propto \mathcal{N}(\Delta(\mathbf{x}_{obs}, \mathbf{x}) | m_\varepsilon, \sigma_\varepsilon^2)$  where the threshold  $\varepsilon$  is replaced by two new parameters  $m_\varepsilon$  and  $\sigma_\varepsilon^2$  that control the similarity of the two data sets. The parameter  $m_\varepsilon$  can be fixed to some expected minimum value of the discrepancy. The likelihood approximation at  $\theta^*$  is then proportional to

$$\tilde{p}(\theta^*) = \int_{-\infty}^{\infty} \mathcal{N}(\Delta | m_\varepsilon, \sigma_\varepsilon^2) \mathcal{N}(\Delta | f(\theta^*), \sigma_n^2) d\Delta \quad (39)$$

$$= \mathcal{N}(f(\boldsymbol{\theta}^*) \mid m_\varepsilon, \sigma_\varepsilon^2 + \sigma_n^2). \quad (40)$$

This approach can be seen as an approximation to the uniform threshold but it could be interpreted also as additional Gaussian measurement (or modelling) error as described by [Wilkinson, 2013].

Proceeding as in Section 3.2 and using the Gaussian identities presented in the appendix of [Rasmussen and Williams, 2006], the expectation and variance can be shown to be

$$\mathbb{E}(\tilde{p}(\boldsymbol{\theta}^*)) = \mathcal{N}(m_\varepsilon \mid m(\boldsymbol{\theta}^*), \sigma_{\varepsilon n}^2 + v^2(\boldsymbol{\theta}^*)), \quad (41)$$

$$\mathbb{V}(\tilde{p}(\boldsymbol{\theta}^*)) = \frac{1}{2\sqrt{\pi\sigma_{\varepsilon n}^2}} \mathcal{N}(m_\varepsilon \mid m(\boldsymbol{\theta}^*), \sigma_{\varepsilon n}^2/2 + v^2(\boldsymbol{\theta}^*)) - \mathcal{N}(m_\varepsilon \mid m(\boldsymbol{\theta}^*), \sigma_{\varepsilon n}^2 + v^2(\boldsymbol{\theta}^*))^2, \quad (42)$$

where  $\sigma_{\varepsilon n}^2 = \sigma_\varepsilon^2 + \sigma_n^2$ . These formulae do not require evaluating special functions and can be used in the place of the Equations (8) and (12). While running the simulation model typically dominates the computational cost, these formulae might be useful in high-dimensional scenarios where one typically needs to evaluate the acquisition function huge number of times. Nevertheless, in this paper we only considered the case with the uniform threshold.